IMPROVING COLLABORATIVE FILTERING FOR NEW-USERS BY SMART OBJECT SELECTION

Arnd Kohrs – Bernard Merialdo
Institut EURECOM
Department of Multimedia Communications
BP 193 – 06904 Sophia-Antipolis – France
{Arnd.Kohrs,Bernard.Merialdo}@eurecom.fr

ABSTRACT

Collaborative filtering is a key technology for recommender systems and personalized services. Collaborative filtering is difficult when few information is known about the user. For new users the collaborative filtering system needs to be trained in order to provide better results. To train a collaborative filtering system a user needs to rate objects.

We propose algorithms for the selection of objects to be rated by new users for a more efficient training of collaborative filtering systems. Our approaches are based on variance and entropy of object rating distributions. We validate the approach in off-line experiments on two different collaborative filtering datasets.
Keywords: Personalization, Collaborative Filtering

1. INTRODUCTION

Modern information-based society evolves rapidly around the Internet. Competing services organize information on Web sites for simultaneous access by millions of users. In general, users are overwhelmed by this new abundance of information. Important issues become apparent: Filtering and recommendation tools are necessary for users to harness the huge amount of information. Additionally, to focus on the users' varying goals and priorities, personalization tools are needed for efficient interaction with Web-sites.

Collaborative filtering is an adequate technology in this context of rich content. Collaborative filtering may be used to filter and recommend arbitrary objects based on opinions (ratings) by users. Collaborative filtering is a relatively new technology and therefore many issues still need to be resolved. One important issue lies in the dependence of a collaborative filtering system's performance on the amount of ratings (quantized opionions) available from users. This lack of data leads to a family of problems comonly known as sparsity issues. One particular issue is the *New-User-Case*: When a new user, one that has not used the system before, uses a collaborative filtering system no personalized (adapted to his taste) recommendations can be provided since no relationship can be determined between known users. For such cases the CF system has to either rely on an alternative fallback algorithm, e.g. using average ratings of the whole population of users, or demand the new user to first to rate some objects in order use the capabilities of collaborative filtering. In such situations there is a trade-off in how much effort users spend in rating objects before they receive any benefits and the gain in precision of predictions of CF due to more information about the user.

In this paper we study this trade-off, and further provide some solutions for making the pre-rating procedure more efficient in terms of either having to rate less objects or getting a better precision out of the CF system with the same amount of effort. We propose methods for smarter selection of objects which should be rated by new users. The proposed methods are experimented on our *Active WebMuseum* project, a web-based virtual museum for art paintings which uses collaborative filtering to provide personalized organization of the content. Data collected in this ongoing trial is used in our research to validate new algorithms in off-line experiments and the most successful methods are used in the implementation of the Web site. To further support our results we conducted also experiments on a movie rating dataset, the EachMovie collaborative filtering dataset[1]. We first describe our application the *Active WebMuseum*. Then after a brief introduction to collaborative filtering, we present our methods for smart object selection, variance and entropy, together with an optimization step followed by results we obtained in experiments to support their validity.

1.1 The Active WebMuseum

In an ideal world a visitor of a museum would enter a museum and then find in the first corridor exactly those items, which he would find most interesting. Given that real museums serve many people at the same time, it is not feasible to rearrange the collection for individual visitors. In contrast, when a museum's art collection is presented through the Web, it becomes feasible to rearrange the collection for each individual visitor. Our Active WebMuseum[2]has a dynamic topology which is adapting to the museum visitor's taste and choices.

The dynamic topology is achieved by dynamic corridors, virtual corridors which contain paintings of a chosen category sorted according to personalized predictions produced by collaborative filtering. See figures 1 and 2 for examples.

While visiting the Active WebMuseum, users may express preferences by giving symbolic ratings to paintings (*excellent*, *good*, *neutral*, *bad*, *terrible* - mapped on a five point scale $1-5$). For paintings which have not been rated by the

[1]The EachMovie dataset also used for this paper was generously provided by Compaq Corporation (formerly Digital Equipment Corporation) (See `http://research.compaq.com/SRC/eachmovie/`).

[2]The Active WebMuseum (accessed through `http://www.eurecom.fr/~kohrs/museum.html`) uses the collection of paintings from the *WebMuseum, Paris*(accessed through `http://metalab.unc.edu/wm/`), which has been created by Nicolas Pioch and contains roughly 1200 paintings by about 170 painters.

Figure 1: In the *Active WebMuseum* the user browses dynamic corridors: When a user has entered a dynamic corridor (in this example a corridor containing paintings by Jackson Pollock), he is presented iconized paintings ordered according to his (possibly predicted by collaborative filtering) preference.

visitor, the ratings are predicted using other users ratings and collaborative filtering technology.

Figure 2: A user may choose a single painting to be viewed in detail. Here, the user may rate the painting in order to express how much he enjoys it. From here the user may visit corridors containing related paintings.

In the following section the underlying collaborative filtering technology is described.

1.2 Collaborative Filtering

Collaborative filtering (CF) systems select objects for a user based on opinions of other users for the same objects. Generally, collaborative filtering systems *do not* rely on content-based information, considering only human judgments on the value of objects. Collaborative filtering systems consider every user as an expert for his own taste, so that personalized recommendations can be provided based on the expertises of taste-related users. Collaborative filtering has been applied to several domains of information: News articles, GroupLens [4]. Music, Ringo [6]. Movies, MovieCritic[3].

Most collaborative filtering systems collect the users' opinions as ratings on a numerical scale, leading to a very sparse matrix $rating(user, object)$ (in short $r_{u,o}$). Collaborative filtering systems then use this rating matrix in order to derive predictions. Several algorithms have been proposed on how to use the rating matrix to predict ratings [2, 6, 1]. For the Active WebMuseum we derived a collaborative filtering algorithm from a commonly used technique, also used in the GroupLens project and in Ringo, which is based on *Pearson* vector correlation. The predictions are weighted sums of other users ratings, and the weights are determined by correlation coefficients between the users' ratings vectors. In the following we describe the underlying formulas in more details.

The task of a collaborative filtering system is to predict the rating of a particular user u for an object o. The system compares the known ratings from user u's ratings with the ratings of all other users, who have rated the considered object o. Then a weighted average of the other users ratings for object o is used as a prediction.

If O_u is set of objects that a user u has rated then we can define the mean rating of user u as:

$$\overline{r_u} = \frac{1}{|O_u|} \sum_{o \in O_u} r_{u,o}$$

When Pearson correlation is used, similarity between users is determined from the correlation of the rating vectors of user u and the other users u':

$$\rho(u,u') = \frac{\sum_{o \in O_u \cap O_{u'}} (r_{u,o} - \overline{r_u})(r_{u',o} - \overline{r_{u'}})}{\sqrt{\left(\sum_{o \in O_u \cap O_{u'}} (r_{u,o} - \overline{r_u})^2\right)\left(\sum_{o \in O_u \cap O_{u'}} (r_{u',o} - \overline{r_{u'}})^2\right)}}$$

It can be noted that$\rho \in [-1, +1]$.

The value of ρ measures the similarity between the two users' rating vectors. A high absolute value signifies high similarity and a low absolute value dissimilarity. The general prediction formula is based on the assumption that the prediction is a weighted average of the other users ratings.

$$p^{CF}(u,o) = \overline{r_u} + k \sum_{u' \in U_o} \rho(u,u')(r_{u',o} - \overline{r_{u'}})$$

$$\text{with} \quad \begin{cases} U_o & : \text{Users, who rated object } o. \\ k & = \frac{1}{\sum_{u' \in U_o} \rho(u,u')} \end{cases}$$

(The factor k normalizes the weights.)

1.3 Selecting objects to be rated by new users

When a user uses a CF system for the first time and no, or very few, objects have been rated by him, the CF system can only perform poorly when compared to the case when many objects have been rated by the user. In [2] the lack of performance of collaborative filtering for users with few ratings is identified as the *New-User-Case*. New users of a collaborative filtering system need to provide some ratings first to obtain personalized results from the CF system. Pennock [5] discusses value of information (VOI) analysis in the context of collaborative filtering, in order to more cost effectively enquire about objects. However, a proposal as to how to proceed is not yet suggested.

We added the option of rating a random sequence of paintings in a batch to the core functionality of the *Active Web-*

[3] `http://www.moviecritic.com`

Museum (see Figure 2 for an illustration of the rating process). New users are requested to use this functionality to *train* the system, and therefore obtain personalized results in the subsequent visit to the museum. While this is necessity, it is however very annoying for visitors of the Active WebMuseum who want to enjoy a personalized dynamic tour focusing on preferred paintings.

Asking for the user to rate random objects is probably not the best way to train a collaborative filtering system. In the following we focus on identifying the most promising objects to be rated in the training phase of new users in contrast to just choosing random sequences of objects. The goals are to demand fewer training ratings from new users and to provide better prediction performance for the same training effort.

2. SMART SELECTION OF OBJECTS

In order to improve upon random selection of objects to rate we try to prioritize objects according to the amount of precision improvement a user gets if he rated the object. This is of course impossible to know before the user actually rates the object. However, in our approach we assume that the users of a CF system are equal with respect to the best sequence of objects to be rated. This assumption permits us to make statistical analyses of the rating data, which is already obtained from prior users, to derive favorable sequences of objects to be rated by new users.

2.1 Theoretical Approach

The task of object selection is to identify the size-limited set S of objects from a given set of potential objects P which when rated will maximize the future prediction performance of the CF system for a target set of objects T. The problem can be generalized in the following way:

$$\underset{S \subset P}{\text{argmax}}(\text{performance}\,(p^{CF}_{/S}(T)))$$

In this article, we study two approaches for finding subsets of objects to be rated by new users, *variance* and *entropy*. Both approaches are based on statistics of the ratings given by other users for the objects in the dataset. The idea is to order all potential objects respective to a statistic taken from their rating distribution. The first approach measures the variance of all ratings each object received.

$$\text{variance}(o) = \frac{\sum_{u \in U_o} (r_{u,o} - \overline{r_o})^2}{|U_o|}$$

The term U_o refers to all users who have rated the object o so far and $\overline{r_o}$ indicates the mean rating assigned to this object. The objects are selected for the set S from the potential set P satisfying the following condition:

$$\forall s \in S, r \in P \backslash S : \text{variance}(s) \leq \text{variance}(r)$$

Our second approach is aimed at answering the question: Which object, when a user's rating for it is known, does reveal the best the users identity? In order to answer this question we consider the random variable U as the identity of a user and the random variable $R(o)$ for the observed rating of a user. Now the entropy$(o) = H(U|R(o))$ of the random variable U under the constraint that $R(o)$ is known can be calculated:

$$H(U|R(o)) = -\sum_{u,r} p(U = u, R(o) = r) \cdot \log(p(U = u|R(o) = r))$$

Here, $p(U = u, R(o) = r)$ is the probability of the event that user u rates object o with r and $p(U = u|R(o) = r)$ is the probability that the u is the user if r is observed as rating for o. The objects are selected in the same way using entropy as before using variance in increasing order.

When several objects are selected at the same time it might occur that similar objects are selected with strongly correlating rating vectors and therefore might not provide as good as a selection as if less correlated objects were chosen. We propose an optimization step which removes the most correlated object of a selected set S and replaces it with the least correlated object of potential objects P:

$$c = \underset{r \in S}{\text{argmax}}(\sum_{s \in S} \text{corr}(r, s)) \quad n = \text{argmin}_{p \in P}(\textstyle\sum_{s \in S} \text{corr}(p, s))$$

$$S := (S \backslash \{c\}) \cup \{n\} \qquad P := (S \backslash \{n\}) \cup \{c\}$$

The optimization step may be repeated a variable number of times.

2.2 Experiments

In order to validate the effectiveness of the sequence selection algorithm of section 2, we set up off-line experiments. In order to compare our results we used random selection (as currently implemented in the Active WebMuseum) as a base-line experiment.

While our target application is the *Active WebMuseum* and we therefore use the limited dataset of ratings collected during the ongoing public trial, we further support results by using the larger dataset of another CF project *EachMovie*. Table 1 lists the dimensions for the datasets in use for the following experiments. The datasets have been reduced from their original size in order to remove users and objects with only few ratings. From the data-sets for some

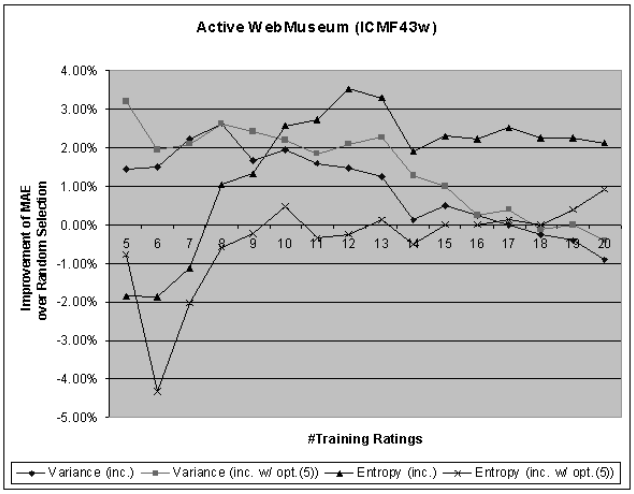
Dataset	users	objects	ratings
Active WebMuseum	468	1116	11500
EachMovie	1315	408	70047

Table 1: Dimensions of the datasets

target users (for whom enough ratings are known) the ratings were divided into a target-set, to be predicted by collaborative filtering, and a *potential* training set. From the potential training set a subset of objects is selected as a training sequence using the algorithm described in the previous section. Then the target-set is predicted using the remaining database of ratings by other users in conjunction with the selected training sequence. The performance of a selected training sequence is measured in prediction precision. The precision is measured as mean absolute prediction error, MAE $= \sum_i \frac{|r_i - p_i|}{n}$, where r_i is a rating in the target-set and p_i is prediction of the rating. The sampling of the selection of training objects is repeated several times, each time initializing the random process differently, and the MAE is averaged over all repetitions. The repetition takes place in order to avoid random effects by the choice of a target-set.

Note, this experimental setup does not fully match the real world application. In the real world application the objects can be chosen from all available objects in the dataset, i.e. 1200 paintings, while in the experimental setup the choice of objects is limited to a potential training set (see above) so that improvements in the real world application are expected to be better than the improvements proven by the results of the experiments.

The measurement comparisons for our *Active WebMuseum* dataset and for the EachMovie dataset are shown in



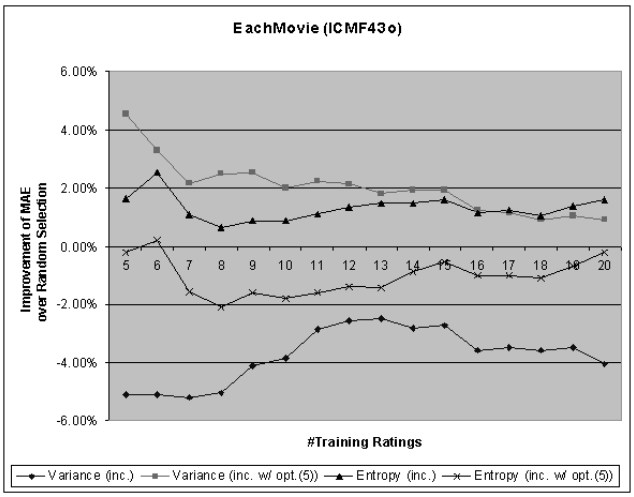


Figure 3: Performance comparison of selection schemes. The experiments are based on a datasets obtained from the Active WebMuseum and on a dataset derived from the EachMovie dataset. A positive percentage indicates a decrease of mean absolute error in comparison with the base selection approach Random selection.

Figure 3. For the EachMovie dataset it can be observed that Entropy consistently outperforms Random selection, while Variance selection consistently underperforms the random selection. For the Active WebMuseum dataset Variance outperforms consistently and Entropy outperforms for more than eight training ratings.

The application of optimization of the selection (5 optimization steps) improves the Variance selection drastically for the EachMovie set. The optimization of the Entropy selection leads to a worse than Random performance for the EachMovie set. Similarly, the optimization degrades the Entropy selection for the Active WebMuseum set. For the Active WebMuseum Variance selection is slightly improved by optimization, though not as drastically as for the EachMovie dataset.

The slight different observations for the two datasets can be explained by the difference in size. In the EachMovie dataset larger sets of rated objects for target users are available than for the Active WebMuseum dataset. However, for both datasets optimized Variance selections consistently outperform Random selections. Also non-optimized Entropy selection seems promising as an alternative selection scheme.

When looking at these results it should be considered that the application of this scheme in a real collaborative filtering system would lead to better results. In the experiment, the potential set of objects is limited to the objects which have been rated by the target users. An even better outperformance of our proposed strategies over random selection should be expected.

3. CONCLUSION

In this paper, we have proposed and evaluated methods to efficiently select objects to be rated by a new user in a collaborative filtering system. Experiments indicate that this smart selection allows to decrease the amount of information required from a user to achieve a given performance, or improves the performance of the prediction for a given number of ratings.

Our methods are directly applicable in many situations of collaborative filtering systems, and adaptations can be considered in others. It is probable that smart selection of training objects will improve the common applicability of collaborative filtering systems.

4. ACKNOWLEDGMENTS:

This research was supported by a regional research grant, and by Eurecom's industrial members: Ascom, Cegetel, France Telecom, Hitachi, ST Microelectronics, Motorola, Swisscom, Texas Instruments, and Thales.

5. REFERENCES

[1] Jack Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Madison, WI, July 1998. Morgan Kaufmann Publisher.

[2] Arnd Kohrs and Bernard Merialdo. Clustering for collaborative filtering applications. In *Computational Intelligence for Modelling, Control & Automation (CIMCA'99), Vienna*, pages 199–204. IOS Press, February 1999.

[3] Arnd Kohrs and Bernard Merialdo. Improving collaborative filtering with multimedia indexing techniques to create user-adapting web sites. In *Proceedings of the 7th ACM Multimedia Conference, Orlando*. ACM, 1999.

[4] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordan, and John Riedl. Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87, March 1997.

[5] D. Pennock and E. Horvitz. Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In *IJCAI Workshop on Machine Learning for Information Filtering, Stockholm*, 1999.

[6] Uprendra Shardanand and Pattie Maes. Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of Human Factors in Computing Systems, CHI '95*, 1995.